One-sample categorical data: the Bayesian approach

Patrick Breheny

September 20

Recap

- In our last lecture, we considered the problem of inferring the survival probability θ of a baby born at 25 weeks gestation, based on the Johns Hopkins study in which 31/39 such babies survived
- In that lecture, we took the long-run frequency interpretation of probability; from this perspective, the survival probability we are interested in is a fixed quantity
- To carry out inference, we constructed a confidence interval according to a procedure guaranteed to contain the true probability of a binomial proportion (at least) 95% of the time

Treating θ as random

- In today's lecture, we'll consider the same problem from the probability-as-uncertainty school of thought
- From this perspective, θ is random – not because it's changing from moment to moment, but because we don't know what it is
- Since θ is now a random quantity, we cannot discuss its "value", but must instead discuss its distribution, $f(\theta)$, which, again, provides a complete description of all the values θ can take on and the probability that it falls within any interval of values

$f(\theta|x)$

- Now, the notion of probability-as-uncertainty is inherently subjective, but we can (and should) at least base our beliefs concerning θ on something objective – namely, the fact that $x = 31$ babies survived
- Thus, what we're really interested in is the distribution of θ based on the data, or more formally, $f(\theta|x)$
- From the perspective of treating θ as random, this conditional probability of the unknown given the data is the focus of all inference, not just in the binomial problem but for any inference of any kind

Bayes rule

- As we have already discussed, it is reasonable to assume that $f(x|\theta)$ is the binomial distribution
- What we need, then, is a way to determine $f(\theta|x)$ based on $f(x|\theta)$
- As you hopefully recall from the lecture on probability, this is exactly the kind of thing you use Bayes rule for
- As we'll see on the next slide, however, there is an added wrinkle here that we haven't seen before – to calculate $f(\theta|x)$, we need to specify $f(\theta)$

Paradigm for Bayesian inference

Letting θ denote an unknown parameter of interest and x observed data, the basic approach to Bayesian inference can be represented as follows:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)},$$

where

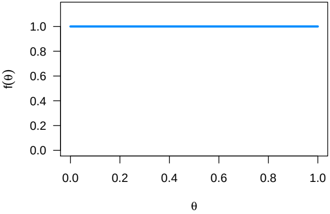
- $f(\theta)$ is the *prior*: Our beliefs about the plausible values of our parameter before seeing any data
- $f(x|\theta)$ is the *likelihood*: The sampling distribution for how the data depends on the unknown parameters
- $f(\theta|x)$ is the *posterior*: Our updated beliefs about the plausible values for our parameter after seeing the data
- $f(x)$ is a normalizing constant typically not of interest

The central role of Bayes rule

- Note that the long-run frequency perspective didn't tell us anything about how to conduct inference – we could construct intervals in any possible way we choose, so long as they had the appropriate coverage probability
- The probability-as-uncertainty perspective, on the other hand, tells us *exactly* how to carry out inference, in every situation: you always use Bayes rule
- Because of this central role of Bayes rule in carrying out all inference according to this perspective, this approach to statistical inference is known as *Bayesian*, as in *Bayesian inference*, *Bayesian statistics*, etc.

Specifying a probability model

- So let's proceed with an analysis of the Johns Hopkins infant survival study from the Bayesian perspective
- In any Bayesian analysis, we need to specify two things:
  - The likelihood $f(x|\theta)$, which in this case is binomial
  - The prior $f(\theta)$
- We'll consider various choices for the prior, but let's start with a *uniform* distribution:



The Beta distribution

- With this model,

$$f(\theta|x) \propto \theta^x (1-\theta)^{n-x}$$

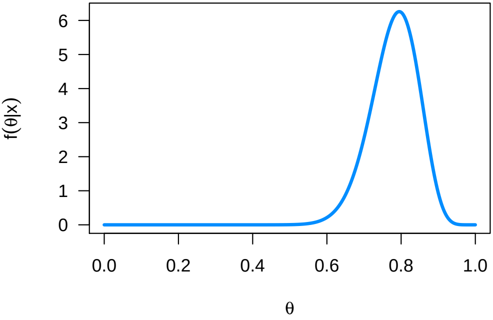
- This falls into a well-known and well-studied family of distributions in statistics known as the beta distribution
- A random variable Y follows a *beta distribution* with shape parameters $\alpha > 0$ and $\beta > 0$ if its pdf is

$$f(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

  over the region $[0, 1]$ and 0 otherwise
- $\Gamma(\cdot)$ is the *Gamma function*, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$

The posterior distribution

Thus, $\theta|x \sim \text{Beta}(x+1, n-x+1)$, which for $n = 39$ and $x = 31$ looks like this:



Posterior intervals

- It would be nice to summarize this distribution with an interval that had, say, a 95% probability of containing θ
- This can be done by evaluating the quantile function (inverse CDF) of the $\text{Beta}(32, 9)$ distribution at the values 0.025 and 0.975
- Although neither the Beta CDF nor its inverse is available in closed form, we can easily calculate these quantities on a computer (which we will do in lab) and determine the 95% posterior interval $[0.644, 0.892]$ for θ

Highest posterior density intervals

This is not the only way to construct a 95% interval:

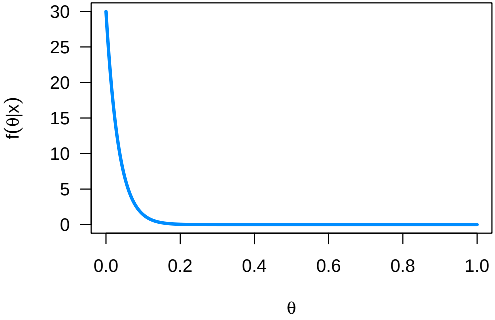
This interval, $[0.654, 0.899]$, is called the *highest posterior density* interval, and is slightly different from the previous interval, which is known as the *central interval*

Confidence intervals vs. Posterior intervals

- So, when looking at what the Johns Hopkins study has to say about survival probabilities for an infant born at 25 weeks gestation, we obtained a confidence interval of $[63.5\%, 90.7\%]$ and a posterior interval of $[65.4\%, 89.9\%]$
- Qualitatively, these two intervals essentially agree, which is reassuring since both approaches seem reasonable and both are "95% intervals"
- Keep in mind, however, that the probabilities these two intervals satisfy are quite different:
  - The 95% for the confidence interval is a statement about $P\{\theta \in [L(X), U(X)]\}$, where θ is fixed and X is random
  - The 95% for the posterior interval is a statement about $P\{\theta \in [L(x), U(x)]\}$, where θ is random and x is fixed because we have conditioned on it

Infant survival: 22 weeks

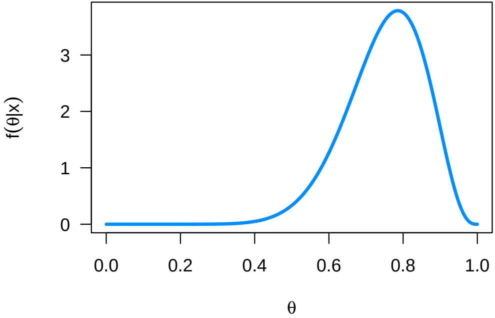
For the same study looking at infant survival at 22 weeks gestation (where 0/29 survived), the posterior looks like:



95% Central interval: $[0.001, 0.116]$
95% HPD interval: $[0.000, 0.095]$

Crossover study

Finally, for our cystic fibrosis crossover trial in which 11/14 patients did better on the drug:



95% Central interval: $[0.519, 0.922]$
95% HPD interval: $[0.544, 0.938]$

Beta prior

- What about other priors besides the uniform?
- Specifically, let's let the prior for θ follow a general Beta distribution:

$$\theta \sim \text{Beta}(\alpha, \beta)$$

  (note that the uniform distribution is a special case of the beta distribution, with $\alpha = \beta = 1$)
- In this case, θ still follows a beta distributon:

$$\theta | y \sim \text{Beta}(y + \alpha, n - y + \beta)$$

Conjugacy

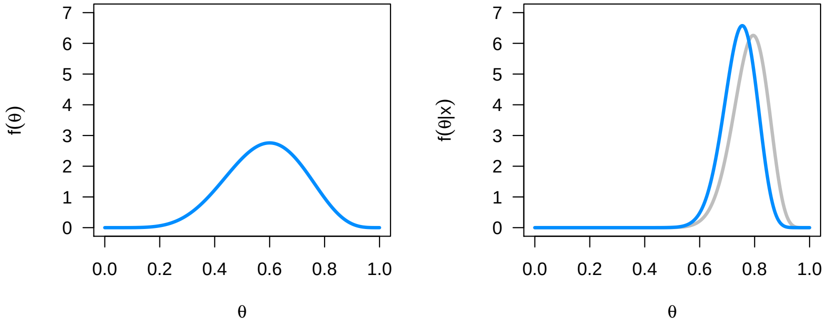
- This phenomenon, in which the posterior distribution has the same parametric form as the prior distribution, is referred to as *conjugacy*
- In this case, the beta distribution is said to be the *conjugate prior* for the binomial likelihood, and is therefore particularly convenient to work with
- We could of course apply Bayes rule and carry out Bayesian inference with any prior, at least in principle, but using conjugate priors makes the math and computing much easier

Informative prior for premature birth data

- Let's suppose that there had been some previous studies that had suggested that the probability of survival for 25 weeks of gestation was around 60%, and that it was rather unlikely to be close to 0% or 100%
- We might propose, in this situation, a $\theta \sim \mathrm{Beta}(7, 5)$ prior
- Note that conjugacy is often helpful when thinking about priors: this is the same as the posterior we would obtain with a uniform prior after seeing 6 successes and 4 failures

25 week survival with $\theta \sim \mathrm{Beta}(7,5)$

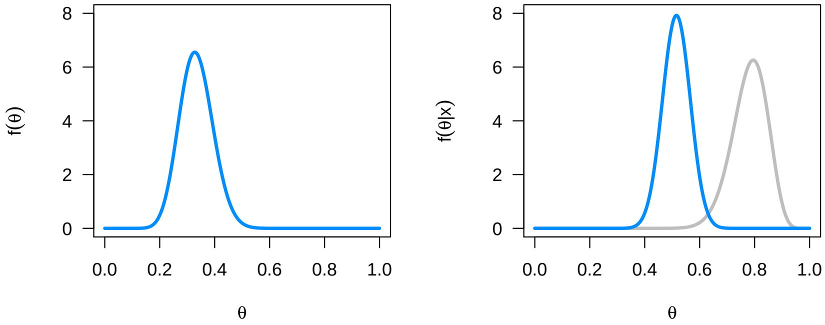
The prior and posterior for $\theta \sim \mathrm{Beta}(7,5)$:



Here, the gray distribution is our previous posterior based on the uniform prior

25 week survival with $\theta \sim \mathrm{Beta}(20, 40)$

Now consider what we happen if we chose a $\mathrm{Beta}(20, 40)$ prior:



Posterior as compromise

- We can see, then, that the posterior distribution represents a compromise between the prior and the likelihood
- Some additional insight into this "compromise" can be gained by considering the posterior mean
- It can be shown that the mean of a beta distribution is $\alpha/(\alpha+\beta)$
- Thus, given a $\theta \sim \mathrm{Beta}(\alpha, \beta)$ prior, the posterior mean is a weighted average of the prior mean and the sample mean:

$$\mathrm{Mean}(\theta|x) = w\frac{x}{n} + (1-w)\frac{\alpha}{\alpha+\beta},$$

where $w = n/(\alpha+\beta+n)$

Sequential updating

- Finally, let's suppose that the data from the infant survival study were collected in two phases: in Phase I, we saw 17/21 infants survive, and in Phase II, we saw 14/18 survive (for a total of 31/39)
- Suppose we started out with a uniform prior, then analyzed the data after Phase I was complete, obtaining $\theta|x_1 \sim \text{Beta}(18, 5)$
- It would be rational to use this as our prior for the analysis of Phase II
- If we do, we would start with a $\text{Beta}(18, 5)$ prior and obtain $\theta|x_2 \sim \text{Beta}(32, 9)$ – exactly the same posterior as before
- Indeed, we could have stopped and analyzed the data after each observation, with each posterior forming the prior for the next analysis; this is known as *sequential updating*

Informative vs. non-informative priors

- Consider our two analyses of the 25-week survival data: one used a uniform prior, while the other attempted to base a prior on previous studies
- Generally speaking, the first prior may be thought of as "non-informative", in the sense that we are just trying to represent a belief that, before seeing any data, all proportions are equally likely
- The other prior, on the other hand, is "informative" in the sense that it is explicitly intended to incorporate external information
- Generally speaking, each type of prior serves different purposes

Informative vs. non-informative priors (cont'd)

- Informative priors are likely more useful for decision making at the individual or organizational level
- Non-informative priors, on the other hand, are useful for communicating results and findings based solely on the data
- To emphasize this point, non-informative priors are sometimes called *reference* priors, as their intent is to provide a universal reference point regardless of actual prior belief
- It is worth noting, however, that the term "non-informative" is somewhat misleading, as all priors contain *some* information

Summary

- Treating θ as a random quantity, Bayesian inference uses Bayes rule to update prior beliefs $f(\theta)$ into posterior beliefs $f(\theta|x)$ based on the data
- To carry out a Bayesian analysis, we must specify both a likelihood $f(x|\theta)$ and a prior $f(\theta)$
- For binomial data, if $\theta \sim \text{Beta}(\alpha, \beta)$,

$$\theta|x \sim \text{Beta}(\alpha + x, \beta + n - x)$$

- There is a basic division among priors between “reference” priors and “informative” priors